UNIVERSITÀ DI PISA

L'Informatica Umanistica tra sociologia della tecnica e linguistica computazionale: come estrarre informazioni

Di

Martina Valeriani

540707

A.a. 2015-2016

LM in InfoUma

Sommario

La premessa

Nell'ambito delle lezioni del Seminario Di Cultura Digitale a.a 2015/2016, le professoresse Marina Caputo e Marion Lamè hanno presentato il loro progetto *Conoscere l'ipertestualità e i suoi dispositivi di comunicazione con Memorata Poetis: una ricerca in Informatica Umanistica nel quadro della sociologia della tecnica. Musisque Deoque, un archivio digitale di poesia latina, dalle origini al rinascimento italiano* è un progetto di ricerca partito alla fine del 2005 con lo scopo di creare un unico database della poesia latina, integrato da apparati critici ed esegetici elettronici.

Ad oggi, le principali collezioni di classici sono state trasferite su supporto digitale ed esistono risorse, per lo più online, che rendono assai celere ogni ricerca lessicale. Nella grandissima parte dei casi, tuttavia, il motore di ricerca si limita a fornire le occorrenze di una chiave all'interno di un testo fisso, 'autoritario'. Si è proposto di superare questa limitazione, permettendo di reperire non solo le forme scelte e riportate dall'edizione di riferimento, ma anche le varianti presentate in apparato.

In tempi più recenti il sito web si è dotato di nuove funzioni. Queste le più significative:

- **epigraphica**: trattamento speciale dei Carmina Latina Epigraphica, con una ricerca per corpora, un incipitario, l'aggiunta di dati quali la provenienza, la datazione, eventuale 'praescriptum' e 'postscriptum' in prosa, ecc.; è stato inoltre avviato un archivio fotografico delle iscrizioni in catalogo;
- **testimoni**: aggiunta, nell'apparato, di una nomenclatura standard dei manoscritti, con la denominazione moderna di città, biblioteca, fondo e segnatura; elenco di autori e opere che condividono lo stesso testimone; collegamento al sito della biblioteca e, se presente, alla versione digitalizzata del manoscritto;
- **ricerca per lemmi**: disponibile nella ricerca avanzata;
- **scansione metrica** di tutte le opere in versi dattilici, fornita dall'applicazione Pedecerto;

- **co-occorrenze**: a partire da un testo sorgente, è esplorato l'intero corpus alla ricerca di somiglianze, verbali o anche ritmiche sopra verbali.
- **Hellenica:** un archivio digitale di poesia greca.

Musisque Deoque è stato, finora, realizzato con finanziamenti del Ministero dell'Università e della Ricerca Scientifica e Tecnologica nell'ambito dei bandi PRIN 2005 e PRIN 2007.[1]

Si tratta di un progetto di digitalizzazione dei testi ideato dal professor Paolo Mastrandrea, in *Strumenti per la filologia digitale*. Lo scopo è quello di catalogare e usufruire di una vasta quantità di testi con una facilità che prima non era possibile nelle biblioteche e arrecando quelli che sono i vantaggi del digitale, come l'accesso a distanza e la collaborazione in linea. Infatti al progetto hanno collaborato a distanza persone in tutta Italia.

La convergenza interdisciplinare tra epigrafia ed epigrafia digitale, filologia e filologia digitale, l'informatica umanistica, l'organizzazione della conoscenza e tecnologie informatiche permettono di avere una conoscenza più globale della trasmissione intertestuale.

Il progetto PRIN e l'Università di Tor Vergata: un caso di archiviazione digitale.

Il PRIN, Progetti di Rilevante Interesse Nazionale, è il titolo della fonte di finanziamento del MIUR che si pone l'obiettivo di promuovere e sviluppare azioni di sistema, favorendo le interazioni tra i vari soggetti del sistema nazionale di ricerca pubblica e altri organismi di ricerca pubblica e privata, e finanziare progetti di ricerca proposti dalle università. Quest'ultimo è il caso del progetto *Memorata Poetis, Memoria poetica e poesia della memoria: Ricorrenze lessicali e tematiche nella*

[1] www.memoratapoetis.it

versificazione epigrafica e nel sistema letterario, strumento digitale concepito per la ricerca di temi e motivi ricorrenti all'interno di un vasto corpus di epigrammi multilingui (greco, latino, italiano antico, arabo).

Si tratta di un esempio di archiviazione digitale che può illustrare cosa voglia dire compiere delle scelte ed escludere delle categorie prediligendone altre.

Il progetto, che è stato ideato da Paolo Mastrandrea[2], non è pubblico: per accedere al sito di *Memorata Poetis*, infatti, è necessaria una password, il quale è diviso in tre cartelle: corpus autore e opera; il dipartimento di lettere e filosofia di Tor Vergata , per quanto riguarda il corpus, si occupa della poesia in volgare delle origini-1375, e delle *Rime* di Boccaccio. L'opera di riferimento viene prima fotocopiata dall'edizione cartacea, poi scannerizzata salvata come documento di Word dove viene analizzata ed eventualmente corretta o modificata e infine salvata come file TXT, pronta per essere inserita nell'archivio digitalizzato. Dopo di che si passa alla tematizzazione: nel caso delle *Rime* ad ogni verso di ciascuna canzone viene inserito un tema, ciascuno dei quali è diviso per categorie con sottocartelle per un totale di migliaia di temi che spesso rendono lungo il procedimento. I problemi relativi alla tematizzazione riguardano la spesso poca congruenza tra un termine all'interno del verso e il tema a cui fa risalire, poiché molti temi sono antiquati ed inoltre in determinate rime o opere non ci sono temi a cui far riferimento e quindi si ricorre all'inserimento di parole chiavi come per esempio "Fiammetta", inserita dal dipartimento di lettere e filosofia di Tor Vergata in riferimento ad alcuni versi delle *Rime.*

Ma il problema più rilevante è quello relativo alla trascrizione delle opere nel *Memorata Poetis*: nell'analizzare le opere di Pier delle Vigne, i collaboratori dell'Università di Tor Vergata si sono trovati di fronte a due varianti diverse della stessa opera da inserire: la lezione di P e quella di V, entrambe valide ma completamente divergenti l'una dall'altra e, di fatto, due testi distinti. Dopo aver

[2] Docente presso l'Università Ca' Foscari di Venezia

discusso a riguardo si è deciso di inserire entrambe le lezioni, poiché sarebbe una perdita rilevante escludere l'una o l'altra, dal momento che entrambe potrebbero essere le più vicine all'ultima volontà dell'autore. La scelta compiuta, in questo caso, è quella di non condannare all'oblio una variante qualora ce ne siano più di una relative alla stessa opera dell'autore in questione.

La scelta però si presenta nell'inserimento delle tipologie letterarie: il progetto è nato per l'archiviazione digitale di forme brevi, epigrammatiche, le quali risultano fin troppo esigue per poter dar vita ad un corpus ricco. Le forme letterarie puramente brevi sono epigrafi ed epigrammi, ma le prime richiederebbero l'intervento di un paleografo e le seconde non sono numerose.

Per questo motivo si è optato per l'inserimento anche dei sonetti, che, però, devono essere considerati il limite entro cui sviluppare il progetto: le canzoni e gli altri generi letterari devono essere necessariamente esclusi, sebbene le *Rime* di Boccaccio presentino numerose canzoni, ormai inserite nell'archivio. Cosa fare? L'episodio del *Musisque deoque*, progetto previo di Mastrandrea e temario attorno al quale si sviluppa *Memorata Poetis*, in cui sono stati inseriti anche i *Carmina Docta* di Catullo nell'ambito degli epigrammi, sebbene non lo siano affatto, ha dimostrato come il *modus operandi* di Mastrandrea sia quello di inserire più opere possibili. Perciò si è deciso di lasciare inseriti tutti i componimenti delle *Rime* ma, poiché ciò comporta un procedere fuori tema venendo meno all'essenza del progetto, di impostare la ricerca semantica solo per componimenti brevi.

Per quanto riguardala trascrizione, si è optato per l'inserimento dei segni diacritici presenti nelle edizioni cartacee: questo perché trattini e puntini sono importanti ai fini della lettura in quanto la loro presenza o meno fa si che venga attribuito al componimento un senso diverso. Per esempio in un componimento di Rinaldo D'Aquino il trattino basso nella parola *nonn_unque* indica divisione netta tra le due parole nella pronuncia, mentre il puntino nella parola *co.leante* indica legame tra le due parole.

Il lavoro di archiviazione digitale comporta dover fare costantemente delle scelte che rispecchino le coordinate del determinato intento con cui si decide di conservare dei testi: in questo caso si tratta di un progetto preciso, che tiene in considerazione tutte le varianti d'autore. Però, il desiderio di non tralasciare nulla, prevale spesso sui limiti entro cui operare. Non è semplice decidere cosa escludere e cosa no, cosa condannare all'oblio e cosa invece ricordare, ma il titolo del progetto forse fornisce già un indizio: *"Memorate"*.

Perché la sociologia della tecnica: le categorie residuali

La sociologia della tecnica indaga il modo in cui la tecnologia prende forma e viene utilizzata, nonché i risvolti sociali che essa assume nell'uso. Come afferma Bruno Latour:

<<È *impossibile trovare un solo ambito in cui gli oggetti esistano senza essere "pieni di persone", così come nessuna società umana può funzionare in modo incorporeo, senza poggiare sul materiale e su tecnologie. Ogni mediazione fra gli uomini, fra persone e cose, si compie utilizzando strumenti tecnici che al loro interno ricomprendono l'insieme di attori eterogenei, almeno in forma simbolica*>>.

Le categorie residuali, che vengono escluse, costituiscono una grave perdita ai fini della conoscenza e della ricchezza culturale, in quanto la scelta della loro esclusione si basa su cosa è utile o serve in un *hinc et nunc* che tralascia l'importanza del ricordo in una prospettiva temporale più ampia, in cui si condannano i futuri lettori o fruitori ad una parziale conoscenza del mondo reale con tutte le sue innumerevoli sfaccettature che non rientrano in un sistema di classificazioni condiviso.

l'approccio ecologico, che si focalizza sulla cooperazione tra tutti gli attanti: marginali, centrali, umani o non umani. Considera le persone, le cose, le informazioni come parte di un insieme che non è possibile scindere, poiché tutte collaborano allo

stesso processo. Cose e persone appartengono simultaneamente a mondi sociali diversi all'interno dei quali gli artefatti mediano l'azione.

Gli oggetti tecnici si presentano come compositi ed eterogenei: rinviano sempre ad un fine, ad un utilizzatore per il quale sono stati progettati, ma al tempo stesso non sono che un'istanza di intermediazione all'interno di una lunga catena che associa umani, prodotti, strumenti, soldi etc. La configurazione stessa dell'oggetto tecnico impone un certo numero di vincoli sulle relazioni che gli attanti stabiliscono tra loro e con l'oggetto in questione. [3]

La tecnologia è possibile proprio perché umani e non umani collaborano tra loro. Il problema è la nuova tendenza a delegare sempre più alle macchine i problemi relativi ai problemi presentati dall'oggetto tecnico.

Nell'ambito della visione ecologica è importante introdurre il concetto di "oggetto liminare", comparso per la prima volta in un articolo accademico di Susan Leigh Star e James. R. Griesemer del 1989 :

<< Boundary objects are objects which are both plastic enough to adapt to local needs and constraints of the several parties employing them, yet robust enough to maintain a common identity across sites. They are weakly structured in common use, and become strongly structured in individual-site use. They may be abstract or concrete. They have different meanings in different social worlds but their structure is common enough to more than one world to make them recognizable, a means of translation. The creation and management of boundary objects is key in developing and maintaining coherence across intersecting social worlds >>. [4]

Gli oggetti liminari sono oggetti abbastanza malleabili da adattarsi ai bisogni locali e ai numerosi set locali che li adoperano, ma robusti abbastanza per mantenere un assetto identitario comune ai vari attori sociali e identico nei vari mondi sociali seppur assumano significati diversi. Ciò risulta molto importante soprattutto perché la tecnologia assume un significato e si differenzia dagli altri oggetti proprio sulla

[3] A. Mattozzi, *Il senso degli oggetti tecnici*,2006, Meletemi ed., Roma.
[4] Bowker G., Star S.L., *Sorting Things Out: Classification and its Consequences*, 1999, The MIT Press, Cambridge, pp. 393.

base dell'uso che se ne fa in diversi contesti sociali: gli oggetti sono classificati come tecnologici in base ai punti di vista e alle finalità degli attori. Proprio per questo motivo risulta palese l'importanza del contesto in cui una tecnologia si diffonde o viene utilizzata: senza contestualizzazione, intesa come associazione del dispositivo con elementi estranei alla sua progettazione, un dispositivo può anche essere perfetto ma non si può realizzare, risulta irreale, astratto. È un concetto fondamentale che permette di capire come la contestualizzazione di un progetto tecnico porti alla formazione di un insieme di relazioni economiche e sociali, a veri e propri saperi nuovi.

Gli oggetti tecnici circolano in mondi sociali diversi, creando un legame che non è solo connesso agli usi, ma anche alla condivisione di saperi e alla creazione di relazioni interstiziali.

Proprio per questo motivo hanno il fine intrinseco della traduzione o naturalizzazione che li conduce ad una stabilizzazione, diversa però da quella che ha come esito il "black boxing" (approccio ANT). Naturalizzazione vuol dire che l'oggetto tecnico tende ad assumere il significato e il ruolo che un determinato mondo sociale decide di attribuirgli. Questo implica che, sebbene ogni dispositivo tecnico, in quanto liminare, sia composto di una parte robusta che gli permette di essere universale almeno a livello ideale, esso viene a configurarsi in maniera diversa in base ai contesti in cui si trova ad essere accolto. Per questo si parla di "traduzione multipla", poiché non può risultare mai univoca. La natura dell'oggetto tecnico è ibrida, così come le reti in cui circola e gli attori al loro interno.

Però, come ci dicono Star e Bowker, le traduzioni multiple sono un lavoro contraddittorio, in cui allo stesso tempo si costruiscono ibridi e si rappresentano essenze: il dispositivo tecnico è frutto di un lungo processo fatto di standard elaborati sulla base di classificazioni frutto di negoziazioni sociali. È all'interno di questo processo che si legittima l'esistenza di dispositivi ibridi attorno ad una infrastruttura informativa fatta di standard (concetti, questi, che analizzeremo successivamente).

Il problema è che gli oggetti sono classificati come tecnologici in base al punto di vista e alle finalità degli attori: la tecnologia assume significato sulla base dell'uso e spesso gli oggetti tecnici veicolano una pluralità di significati. Per gruppi di utilizzatori diversi, allo stesso oggetto materiale corrispondono due artefatti diversi, come ci spiega Bijker attraverso la sua analisi dell'Ordinary, la bicicletta a ruota alta, che per i non utilizzatori risultava essere difficile da guidare e rischiosa in quanto facilmente ribaltabile frontalmente, mentre per il gruppo egli utilizzatori risultava essere un "biciclo macho", ancora rischioso ma per questo entusiasmante, e comunque efficace.

Questa fase di coesistenza di diverse interpretazioni degli oggetti, viene definita da Bijker, flessibilità interpretativa, in cui la compresenza di diverse interpretazioni crea conflitti, controversie e negoziazioni. Per questi motivi ogni vicenda tecnologica ha molteplici esiti in base alle interpretazioni che ne hanno indirizzato lo sviluppo.

Nel processo di traduzione si definiscono i ruoli e le mansioni da attribuire ad attori ed oggetti tecnici, spesso già insiti nello "script" (sceneggiatura) del progettista di un dato dispositivo tecnico, che presenta l'utilizzatore ideale a cui tale dispositivo è indirizzato. Il problema è che molto spesso utilizzatore ideale e utilizzatore reale non coincidono. L'utilizzo ad opera di progettisti e dei costruttori di dispositivi tecnici per raggiungere determinati attori o attribuire ruoli specifici non sempre risulta possibile. Non sempre gli attori sono disposti ad accettare il ruolo proposto da una determinata tecnologia, e ciò mette in crisi il successo stesso di un dispositivo tecnico. La descrizione è il processo che ci permetterà di capire il ruolo sociale e politico di un determinato oggetto tecnico.

Come studiare ciò che è invisibile: standard e classificazioni.

Standard, classificazioni e le infrastrutture informative a cui danno vita, costituiscono la parte invisibile della struttura delle tecnologie: si compongono di dati e meta dati

che permettono alle informazioni di acquisire assetti universali e comuni accessibili in egual misura ai diversi gruppi sociali.

Una classificazione è un segmento spaziale, temporale o spazio- temporale del mondo: un sistema di classificazione è come una "serie di scatole" dentro le quali possono essere messe le cose per poi utilizzarle in ambiti specifici. I sistemi di classificazione sono unici per ogni categoria ed effettuano una totale copertura del mondo che descrivono, in cui tutto deve essere ridotto a dati, ad informazioni conoscibili ed etichettabili. Classificazioni e standard sono strettamente collegati tra loro ma non identici: lo standard è una componente, è un *modus operandi* di classificare il mondo: uno standard consiste in una serie di regole concordate per la produzione di oggetti e per questo abbraccia più di una comunità di pratica . Il termine "comunità di pratica", elaborato da Lave e Wenger, è un altro modo di denominare un mondo sociale, così chiamato invece da Strauss, e consiste in un insieme di legami sociali tra le persone della stessa comunità, che sono interconnesse tra loro e tra gli altri membri appartenenti a diversi mondi sociali attraverso attività di collaborazione.[5]

Una infrastruttura informativa è essenzialmente un insieme di classificazioni e per questo implica il collegamento tra esperienze acquisite in un determinato momento e luogo e quelle acquisite in un altro momento e luogo, tramite rappresentazioni di qualche tipo. L'infrastruttura informativa è ciò che permette la trasmissione di informazioni tra diversi mondi sociali, ed è quindi ciò che garantisce lo spostamento di informazione da un contesto ad un altro per fornire a tutti un mezzo di accesso ad essa attraverso spazio e tempo: questo cambiamento di contesto implica l'eterogeneità dell'informazione stessa ed è ciò che garantisce ogni infrastruttura informativa. L'International Classification of Desease (ICD), per esempio, fa spostare l'informazione in tutto il mondo attraverso i decenni, e attraverso molteplici conflitti di conoscenze mediche e sistemi di pratica; Internet , ad oggi, è la più grande

[5] Bowker G. e Star S. L., Sorting Things Out: Classification and its Consequences, 1999, The MIT Press, Cambridge.

e potente infrastruttura informativa che esista, con i suoi protocolli e standard tramite cui veicola e regola le informazioni provenienti da tutto il mondo ogni millesimo di secondo, garantendo l'accesso a queste fonti a quasi tutta la popolazione mondiale dotata di un dispositivo capace di collegarsi.

Gli standard vengono impiegati per garantire la collaborazione all'interno dei diversi mondi sociali nonostante la distanza e metri di giudizio eterogenei: esistono standard che collegano il computer alla rete telefonica per inviare segnali da una rete all'altra. Classificazioni e standard sono due facce della stessa medaglia, in quanto le classificazioni possono diventare standardizzate o meno e se non lo diventano, rimangono limitate ad una sola comunità locale o comunque di durata limitata. Allo stesso tempo ogni standard di successo impone un sistema di classificazione.

Classificazioni e standard: la loro essenza di oggetti liminari

Poiché abitano più comunità di pratica, soddisfano i requisiti informazionali di ognuno di essi e so.no in grado di viaggiare attraverso i margini dei diversi mondi sociali pur mantenendo una sorta di costante e universale identità, l'essenza di classificazioni e standard non è altro che quella di oggetto liminare, marginale.

L'informazione risiede in più di un contesto e sappiamo che qualcosa è in relazione a ciò che non è: per esempio il silenzio permette di sentire le note musicali. I diversi contesti dell'informazione devono essere ricollegati attraverso una sorta di giudizio di equivalenza e comparabilità: l'informazione è davvero informazione solo quando ci sono molteplici interpretazioni. Il modello ecologico proposto aggiunge il fatto che le persone siano interpreti attivi dell'informazione che abita in contesti d'uso e pratica diversi: questa molteplicità è fondamentale, non accidentale né incidentale.[6]

[6] Bowker G. e Star S.L., *Sorting Things Out: Classification and Its Consequences*, The MIT Press, Cambridge, 1999.

Consideriamo per esempio il progetto di un sistema informatico: Elvil Back ha studiato l'evoluzione di un tale sistema analizzando come due autori, che si trovavano in posti differenti, potessero scrivere insieme un articolo accademico insieme facendo uso dei computer. Il lavoro che stavano facendo e il modo in cui lo conducevano era inseparabile dal contesto a dalla cultura a cui entrambi facevano riferimento. Per far sì che l'intero sistema funzionasse si sono dovuti destreggiare tra fusi orari e avere sensibilità circa le diverse parti della pratica lavorativa, come interpretare e far confluire le frasi l'uno dell'altro e far fronte agli aspetti tecnici relativi alla descrizione di un software e di un hardware. In sostanza, dovettero dar vita ad un contesto condiviso nel quale l'informazione potesse acquisire un senso comune ad entrambi. E' ciò che Suchman e Trigg chiamano "integrazione artificiosa" di vincoli locali, dove si accoglie l'applicazione di standard e la rappresentazione dell'informazione.[7]

Se sia le persone che gli oggetti dell'informazione abitano molteplici contesti e lo scopo centrale di un'infrastruttura informativa è trasmettere informazione attraverso i vari contesti, allora una rappresentazione è una sorta di "sentiero" che include tutto ciò che popola questi contesti. I requisiti fondamentali per una comprensione ecologica del percorso di ri-rappresentazione consiste nell'analizzare e capire:

- come gli oggetti possano abitare molteplici contesti allo stesso tempo ed avere simultaneamente significati locali ma condivisi;

- come le persone che vivono in una comunità e traggono i loro significati dalla sfera di persone ed oggetti del loro mondo sociale, possano comunicare con coloro che invece abitano un mondo sociale diverso;

- come sia possibile modellare l'informazione ecologica di persone e cose attraverso più comunità.

[7] Bowker G. e Star S.L., *Sorting Things Out: Classification and Its Consequences*, The MIT Press, Cambridge, 1999.

La risposta è costituita proprio da standard e classificazioni, le quali si caratterizzano come potenti tecnologie, poiché incorporate nelle infrastrutture lavorative diventano relativamente invisibili senza perdere nulla del loro potere. Il modello ecologico permette di capire come ogni classificazione sia inserita e abbia significato in uno o più contesti: esse sono interconnesse e a loro volta connettono i diversi attori sociali impegnati in un lavoro di collaborazione in cui i diversi significati attribuiti al medesimo oggetto risultano tutti allo stesso modo rilevanti, secondo il concetto di traduzione multipla, alla cui base vi è l'oggetto liminare**,** cioè classificazioni e standard.

Naturalizzazioni multiple: il problema dell'archiviazione e della conservazione di dati nel tempo

L'ubiquità è la prima caratteristica che definisce classificazioni e standard, le quali saturano il nostro ambiente e ogni giorno a migliaia vengono utilizzate per trasmettere dati e informazioni: per questo si caratterizzano come "materiali" ma anche "trasparenti", "ideali", in quanto sono un insieme di elementi fisici, come moduli cartacei, dispositivi tecnici, spine etc. e disposizioni convenzionali come velocità e dimensione. Così quando alcuni programmatori del codice Java, si muovono all'interno di vincoli convenzionali e allo stesso tempo utilizzano strumenti materiali come oggetti tecnici, creano documenti sul desktop e consultano manuali per standard e altre informazioni.

L'invisibilità è un'altra caratteristica degli standard e delle infrastrutture informative, e si presenta come il risultato del processo di " naturalizzazione", per cui risultano ormai talmente parte della nostra vita quotidiana da essere date per scontati, quasi come accadeva per gli oggetti tecnici ritenuti delle scatole nere: gli interruttori della luce, per esempio, sono parti ordinarie della vita moderna e la maggior parte di coloro che vivono nel mondo industrializzato conoscono lampadine ed elettricità anche se

non dovessero venirne a contatto, perché si tratta di conoscenze standardizzare e naturalizzate talmente tanto da risultare ormai invisibili ai nostri occhi seppur sempre presenti nella memoria collettiva della comunità di pratica. La naturalizzazione è di per sé multipla perché la molteplicità è alla base di ogni processo socio tecnico per quanto riguarda il modello ecologico e non può essere eliminata in nessun modo: quindi avviene a livello dei diversi mondi sociali , i quali inglobano ciascuno in modo diverso lo stesso oggetto tecnico[8] a cui si fa riferimento.

Purtroppo, per ogni naturalizzazione e standardizzazione che si effettua per garantire collaborazione e interconnessione si vengono a creare categorie "residuali", "marginali", in quanto non rientrano negli schemi standardizzati delle infrastrutture informative. Una categoria residuale è qualcosa che semplicemente non può essere contenuta in un "solo contenitore": per esempio sono ritenute marginali le persone che appartengono a più di una comunità di pratica o gruppo sociale o tutti quegli oggetti tecnici che risultano diversi ma allo stesso tempo universali per i diversi attori sociali. Tutti siamo membri di molteplici mondi sociali, ma come possiamo naturalizzare diversamente lo stesso oggetto se naturalizzazione significa lasciare fuori altri mondi sociali o altre informazioni? Proprio la naturalizzazione multipla, che salvaguardia la ricchezza dei diversi punti di vista, allo stesso tempo crea categorie residuali che coesistono in più mondi sociali, conoscenze tacite che spesso non vengono tenute in considerazione. oggetti che si oppongono alla naturalizzazione.

Le conoscenze tacite sono quei saperi informali che rimangono inespressi e non possono essere rappresentati in termini discreti e quindi standardizzati: si tratta per lo più di azioni umane e per questo risultano non codificabili e impossibilitati quindi a circolare all'interno dei diversi mondi sociali; allo stesso tempo però risultano fondamentali per la riuscita di un dispositivo tecnico, come nel caso del fallimento del programma nucleare iracheno, individuato proprio grazie all'esperienza dei

[8] Con oggetto tecnico, nel modello ecologico, ci si riferisce ad oggetti materiali quando a dati, standard ed informazioni.

progettisti e a esperimenti, test di prova, testimonianze e scambi di opinione. Classificazioni, standard, infrastrutture informative e naturalizzazioni multiple sono finestre sul mondo che ci permettono di venire a conoscenza di informazioni e di poter manipolarle, ma necessariamente operano in modi selettivi e restrittivi: Bowker e Star le hanno studiate per mettere in luce sia i lati positivi che quelli negativi, mostrando come valorizzino qualche punto di vista e ne passino sotto silenzio altri, e ciò non risulta affatto diverso da ciò che accadeva nei modelli precedenti. Ciò che rimane all'esterno viene dimenticato e così si viene a creare quello che Lotman chiama sistema della "dimenticanza", per cui gli oggetti dimenticati vanno a confluire nella categoria dei non testi o addirittura eliminati fisicamente e questa mansione verrà sempre più affidata alle macchine e ai dispositivi, i quali sempre più decidono quali dati conservare e archiviare e quali no.

Gli studi di Bowker e Star sono nati con lo scopo non solo di rendere visibile ciò che, senza analisi approfondite, risulta trasparente, invisibile all'interno di un processo socio tecnico e che invece è parte fondamentale al suo interno, ma anche di rendere le persone più consapevoli della persistenza di grandi quantità di dati e di come questi vengano gestiti dalle piattaforme informatiche, dai social network, dalle organizzazioni e da noi stessi, quali siano i meccanismi per cui si decida quale informazioni ritenere importanti e quali no. Ogni informazione ed ogni punto di vista sono importanti e standardizzazioni e classificazioni devono esistere per garantire la conoscenza di tutto ciò che ci circonda, senza creare categorie residuali o "mostri" per evitare che la memoria originale venga ad essere sostituita da una "artificiale" , che il mondo esplorato scientificamente diventi sempre più strettamente legato al mondo che può essere rappresentato nelle teorie di qualcuno e nel suo database e che un tale mondo venga sempre più riconosciuto come il vero mondo.

Classificazioni e standard sono potenti tecnologie che mediano il nostro rapporto col sociale, anzi lo regolano quasi completamente, e conoscerne i lati postivi come quelli negativi è importante per rendere consapevoli e capire come utilizzarli al meglio. Soprattutto nei lavori di archiviazione digitale è bene tenere a mente gli studi di

Bowker e Star, poiché è soprattutto in questo campo che spesso si effettuano scelte relative a quali dati prendere in considerazione e quali no.

Annotazione linguistica: un altro caso di sociologia della tecnica nell'ambito della linguistica computazionale

Durante il corso di Linguistica computazionale con la professoressa Simonetta Montemagni, ci è stato chiesto di partecipare a un progetto di annotazione linguistica di un corpus giornalistico u più livelli: morfo-sintattico e sintattico con revisione manuale sulla base di una automatizzata. Annotare a livello morfo-sintattico vuol dire assegnare a ogni token la categoria grammaticale a cui fa riferimento in un contesto dato, come sostantivo, aggettivo, verbo. Si tratta di ricavare informazione dai dati linguistici, una informazione di tipo categoriale che si esprime attraverso etichette, tagset par l'appunto. L'informazione relativa alla categoria grammaticale può essere arricchita da ulteriori specificazioni morfologiche o features, come genere, numero, persona, tempo, modo. Le stesse categorie grammaticali di base possono essere articolate in sottoclassi: i sostantivi si distinguono in nomi comuni e nomi propri, gli aggettivi in possessivi e determinativi, i pronomi in personali, clitici, ecc.[9] Il lavoro di revisione manuale dell'annotazione morfo-sintattica è stata effettuata a partire da un corpus precedentemente annotato automaticamente, costituito da 5293 token e facente parte di un corpus più grande di articoli di giornale. Come tagset di riferimento per l'assegnazione delle parti del discorso abbiamo consultato l'Universal Dependencies POS tags[10] e l'ISST-TANL Tagsets[11]. Invece per la classificazione delle features abbiamo seguito le denominazioni presenti in Universal Dependencies alla voce features[12].

[9] Lenci A., Montemagni S., Pirrelli V. (2015), Testo e computer, elementi di linguistica computazionale, Carocci Ed., Roma
[10] http://universaldependencies.org/it/pos/index.html
[11] http://www.italianlp.it/docs/ISST-TANL-POStagset.pdf
[12] http://universaldependencies.org/it/feat/index.html

La revisione gold in questo caso è stata fondamentale per correggere tutte queste tipologie di errori che altrimenti a livello automatico non sarebbe stato possibile disambiguare. Questo perché gli algoritmi non sempre sono in grado di riconoscere la funzione delle parti del discorso in base al contesto dato, competenza che invece è propria di un annotatore in base al sistema linguistico a cui fa riferimento. È il caso di "lo scolo" in cui lo era stato annotato come articolo e di conseguenza scolo veniva classificato come sostantivo, quando invece si trattava di un pronome clitico che si riferiva ad un complemento oggetto antecedente retto dal verbo in questione ("lo scolo" si riferisce a "scolo il coniglio"). Tuttavia, l'annotazione automatica è in grado di riconoscere parole sconosciute causate da errori di battitura o parole straniere, e di saper comunque assegnare loro il giusto tagset: per esempio nel corpus la parola *terrirori* (territori) è stata riconosciuta come nome maschile plurale.

La seconda fase del progetto è relativa all'analisi sintattica gold del corpus annotato automaticamente, basato sulla annotazione morfosintattica revisionata manualmente nella prima fase. L'annotazione sintattica può essere di due tipi: a costituenti o a dipendenze, come nel caso in analisi. La prima si basa sul riconoscimento di costituenti sintattici, quali sintagma verbale e nominale, e le loro relazioni gerarchiche; la seconda descrive la frase attraverso relazioni binarie (di tipo grammaticale) di dipendenza tra parole, come soggetto, oggetto, verbo. Infatti, l'informazione che viene estratta tramite questo tipo di annotazione è di tipo relazionale. La denominazione comunemente utilizzata per ogni corpus annotato a livello sintattico è treebank, cioè banche di alberi sintattici: la treebank di riferimento utilizzata per l'italiano è l'Italian Stanford Dependency Treebank (ISDT)10, realizzata per il dependency parsing shared task di Evalita (Bosco et al. 2014). ISDT presenta una serie di esempi annotati in accordo con gli schemi di dipendenza di Stanford, ottenuti tramite un processo di conversione semi-automatica. Ciò a partire dal MIDT (Merged Italian Dependency Treebank), frutto della convergenza tra due

treebank italiane diverse tra loro: la TUT (Turin University Treebank) e la ISST TANL (Italian Syntactic - Semantic Treebank).

Essendo l'annotazione linguistica un processo incrementale, risulta importante il collegamento tra analisi morfo-sintattica e analisi sintattica: la prima influenza la seconda, soprattutto a livello di errori.

Annotare un testo, a qualsiasi tipo di livello, significa estrarre informazione linguistica per renderla accessibile e *machine readable*. Si tratta, quindi, di un lavoro importante a livello testuale ma anche computazionale, che richiede dedizione, comprensione, tempo e accuratezza. Per questo è necessaria la collaborazione tra strumenti di analisi automatizzati e le conoscenze di riferimento degli annotatori umani: gli algoritmi da soli non possono ovviare a quegli errori invece disambiguati nella revisione manuale e allo stesso tempo senza strumenti automatici il lavoro di annotazione richiederebbe un tempo e un dispendio di energie troppo elevato a cui la revisione manuale non potrebbe far fronte. Questo anche in termini di qualità, accuratezza ed efficienza.

Conclusioni

Lo scopo della trattazione presentata è cercare di comprendere e far comprendere quanto sia importante, nell'ambito di qualsiasi disciplina, l'interoperabilità tra macchine e umani: di qui l'importanza dell'informatica Umanistica, vista come l'utilizzo di strumenti informatici con la consapevolezza delle loro potenzialità e dell'arricchimento che possono apportare alle materie umanistiche e viceversa. Si tratta di utilizzare gli strumenti tecnologici in maniera critica per una più approfondita analisi delle *Humanities* ricordando che "*la macchina è frutto del pensiero e del lavoro umano"* (Karl Marx). Quindi il lavoro dell'uomo che sta dietro ogni parser testuale semi-automatico può essere implementato e migliorato grazie al confronto tra annotazioni automatiche e annotazioni gold.

Il digitale permette l'interconnessione tra discipline diverse e rispecchia un modo di vedere tipico dell'approccio ecologico all'interno della *Tecnologies Science:* tutte le conoscenze sono interconnesse, così come la tecnologia con la società, la politica e l'economia. È tutto un unico grande sistema, caratterizzato da diverse forze legate tra loro e reciprocamente influenzabili. Come studiare ciò che invisibile? Le classificazioni, i dati, gli standard permeano la nostra vita quotidiana e sono, nella realtà sociale, economica e politica attuale e futura, la vera essenza delle tecnologie. Ogni volta che un utente naviga su Internet entra in contatto con numerosissimi standard, informazioni classificate e dati in quantità abnormi. Il tutto risulta trasparente apparentemente, ma studiando attentamente i meccanismi con cui le nuove tecnologie operano ci si può abituare a rendere visibile ciò che prima non lo era, a scorgere il positivo e il negativo di ciascun processo.

Compiere un'analisi ecologica vuol dire aver capito questo ed essere in grado di scorgere le strutture invisibili di ciò che ci circonda

<<What is Digital Humanities? The intersection between Arts&Humanities disciplines and technology>>

David Beavan

Bibliografia

A. Mattozzi, *Il senso degli oggetti tecnici*,2006, Meletemi ed., Roma.

Bosco, C., Montemagni, S., & Simi, M. (2013, August). *Converting italian treebanks: Towards an italian stanford dependency treebank*. In Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse

Bowker G., Star S.L., *Sorting Things Out: Classification and its Consequences*, 1999, The MIT Press, Cambridge.

Caputo M., Lamè M., *Conoscere l'intertestualità e i suoi dispositivi di comunicazione con MP: una ricerca in Informatica Umanistica,* seminario di cultura digitale, 20/04/2016.

Lenci A., Montemagni S., Pirrelli V. (2015), *Testo e computer, elementi di linguistica computazionale*, Carocci Ed., Roma.

Nivre, J. (2015, April). *Towards a universal grammar for natural language processing*. In International Conference on Intelligent Text Processing and Computational Linguistics (pp. 3 16). Springer International Publishing.

Nivre, J. (2005). *Two notions of parsing. Inquiries into Words, Constraints and Contexts*, 106.

Nivre J., (2005), *Two Strategies for Text Parsing, Edie Tor and Ed Itor* (eds.).

Star S.L, Griesemer J.R (1989), *Institutional Ecology, Translations And Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology*, articolo accademico.

Valeriani M., (2014), *Il modello ecologico nella sociologia della tecnica: studiare ciò che è invisibile,* Tesi di laurea triennale in Lettere e comunicazione multimediale presso l'Università di Tor Vergata, Roma.